DOCUMENT RESUME

ED 453 250 TM 032 787

AUTHOR Miller-Whitehead, Marie
TITLE Evaluation of a Kindergarten Diagnostic Assessment Instrument by Gender and Ethnicity.
PUB DATE 2001-04-13
NOTE 67p.; Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE EDRS Price MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Diagnostic Tests; Educational Planning; *Ethnicity; Kindergarten; *Kindergarten Children; Preschool Children; Primary Education; Public Schools; Reliability; *School Readiness; Screening Tests; *Sex Differences; Student Placement; Test Format; *Test Use; Validity

ABSTRACT

Prekindergarten students (n=1,137) in an ethnically and socioeconomically diverse urban public school took a kindergarten diagnostic assessment in September prior to entry into the regular kindergarten program. The assessment was a pilot version of a longer instrument that had been used at the school for several years. Results were used for placement, program planning (class size and multiage grouping), remediation, instructional improvement, and counseling with parents. The results of the diagnostic assessment were evaluated to determine differences in student readiness for kindergarten by gender, ethnicity, socioeconomic status, and English language proficiency. The objective of this analysis was to determine the reliability and validity of the newer, abbreviated version of the kindergarten assessment and its usefulness in making decisions about individuals and groups of students. This paper provides examples of item domain goals and objectives, identifies items that could be useful to equate with other prekindergarten assessments, examines differences in skill attainment for male and female children from ethnic minority groups not identified as Limited English Proficient (LEP) or English Language Learners (ELL) and provides comparisons of several approaches for psychometric evaluation of test items including the transformed item difficulty index, normal curve equivalents, item characteristic curves, and two-parameter Rasch modeling. Implications for assessment and instruction of children from different cultural and ethnic backgrounds are discussed as well as implications for addressing differences in skills attainment for male and female children as they progress through school. Three appendixes contain a correlation matrix for three assessments, a table of sample norms, and a table of normal curve equivalents. (Contains 8 tables, 12 figures, and 58 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 453 250

Running head: KINDERGARTEN DIAGNOSTIC ASSESSMENT

Evaluation of a Kindergarten Diagnostic Assessment Instrument

by Gender and Ethnicity

Marie Miller-Whitehead
Tennessee Valley Educators for Excellence
TVEE.ORG

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

M. Miller-Whitehead

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM032787

Paper Presented
At the Annual Meeting
Of the
American Educational Research Association
Seattle, WA
April, 2001

BEST COPY AVAILABLE

Evaluation of a Kindergarten Diagnostic Assessment Instrument

Abstract

Prekindergarten students in an ethnically and socioeconomically diverse urban public school system took a kindergarten diagnostic assessment in September prior to entry into the regular kindergarten school program. The assessment was a pilot version of a longer instrument that had been used for several years previously in the school. Results were used for placement, program planning (class size and multiage grouping), remediation, instructional improvement, and counseling with parents. The results of the diagnostic assessment were evaluated to determine differences in student readiness for kindergarten by gender, ethnicity, socioeconomic status, and English language proficiency. The objective of this analysis was to determine reliability and validity of the newer, abbreviated version of the kindergarten assessment and its usefulness for making decisions about individual students and groups of students. This paper provides examples of item domain goals and objectives, identifies items that could be useful to equate with other prekindergarten assessments, examines differences in skill attainment for male and female children and for children from ethnic minority groups not identified as Limited English Proficient (LEP) or ELL (English Language Learners), and provides comparisons of several approaches for psychometric evaluation of test items including the transformed item difficulty index (TID), NCEs, item characteristic curves (ICC), and two-parameter Rasch modeling.

Implications of findings for assessment and instruction of children from different cultural and ethnic backgrounds are discussed as well as implications for addressing differences in skills attainment for male and female children as they progress through school.

Evaluation of a Kindergarten Diagnostic Assessment Instrument
by Gender and Ethnicity

Purpose[1]

Prekindergarten students in an ethnically and socioeconomically diverse urban public school system took a kindergarten diagnostic assessment in September prior to entry into the regular kindergarten school program. The performance assessment was a pilot version of a longer instrument that had been used for several years previously in the school system and was administered to all prekindergarten students except those excluded by IEP. Results were used for placement, program planning (class size and multiage grouping), remediation, instructional improvement, and counseling with parents (Achilles, 1999; National Education Goals Panel, 1998; Nye, 1993; Slavin & Fashola, 1998). The results of the diagnostic assessment were evaluated to determine differences in student readiness for kindergarten by gender, ethnicity, socioeconomic status, and English language proficiency. The objective of this item analysis was to determine reliability and validity of the newer, abbreviated version of the kindergarten assessment and its usefulness for making decisions about individual students and groups of students. This paper provides comparisons of several approaches for evaluation of test items. Since the items on the Kindergarten Diagnostic Assessment reflected comprehension and skills that students would be expected to demonstrate during the kindergarten year the examination of the assessment served primarily as a safeguard against including items that might be misunderstood by students of different national origins and to monitor progress of the diverse

---

[1]This is a baseline report for a project supported in part by an NAEP training grant awarded through the National Center for Education Statistics, Washington, DC.

student population entering the school system, many of whose families spoke languages other than English in the home. Students who remained in the system would take the Stanford Achievement Test (SAT9) in grade 3. Appendix A provides a correlation matrix for the items of the Kindergarten Diagnostic Assessment with NAEP and Stanford Achievement Test (9) objectives. Thus the information provided by results of this assessment might provide some baseline data for evaluation of student progress and growth in knowledge and skills based on gain between kindergarten and that demonstrated on later standardized tests such as the SAT9 and the NAEP. Given the diverse population of the school district a primary concern was to provide an educational experience that would assure acceptable levels of gain mandated by various state and federal programs for students in the subgroups of interest: females, minorities, low-income, LEP or ELL, and IEP students[2]. Thus the items of the kindergarten diagnostic instrument should reflect the most common curricular goals and objectives for prekindergarten students and the items of the assessment should be capable of discriminating a broad range of student ability without bias.

Of particular interest for this paper was examination of gender differences on certain items between prekindergarten males and females. The commonly held view is that females outperform males on verbal ability tests and test items and males outperform females on tests related to spacial, mathematical, and science ability. Therefore the present study sought to identify the extent to which prekindergarten male and female performance on test items measuring those skills was consistent with later high school and college level test data (Childs, 1990, Cole, 1997). The popular conception is that "girls have greater verbal ability, boys excel

---

[2]Title I of the ESEA (Elementary and Secondary Education Act), PL 103-382 (IASA, Title VII, E, §7501(8)), U. S. Civil Rights Acts of 1964, Title VI).

in visual-spacial ability and in mathematics, and males are more aggressive" (Cole, p. 6). The results of these comparisons would be expected to provide valuable information for those planning programs to encourage female and minority students to prepare for careers in science and mathematics as well as for those planning programs to encourage males interested in the liberal arts.

Review of Related Literature

There has been extensive research devoted to examination of test item functioning from a variety of perspectives. Testing companies, educational researchers, and stakeholders involved in equity and fair testing issues have devised methods of assuring that their tests maintain rigorous standards without discriminating against subpopulations of tested students (Bracey & Rudner, 1992; Camilli & Shepard, 1994; Hambleton, Swaminathan, & Rogers, 1991; Linn & Harnisch, 1981; Lord, 1980; Osterlind, 1983; ; Rudner, 1994; Rudner, Getson, & Knight, 1980; Wright, 1977). While the purpose of this paper is not to provide an extensive discussion of test bias but rather to examine ways in which schools determine how best to serve the instructional needs of all their students, it is nevertheless helpful to define key terms at the outset. Test bias can be defined as "invalidity or systemic error in how a test measures for members of a particular group...it creates a distortion in test results for members of a particular group" (Camilli & Shepard, p. 8). Another definition of test bias is where "two groups are equal on the criterion but vastly different on the test" (p. 10). Thorndike defined bias as "performance fair" and Cleary as a function of "test fair" based on different interpretations of regression test fairness models. Educational practitioners including school system administrators, school principals, and teacher teams look to the effective schools research to assure equality of opportunity for identified subpopulations of students who enter or exit their schools. Another widely accepted definition

of differential item functioning is "an item shows DIF if individuals having the same ability but from different groups do not have the same probability of getting the item right" (Hambleton et al., p. 110). Effective schools are those that "can demonstrate the joint presence of quality and equity" with "no differences in the distribution of that achievement among the major subsets of the student population" (Lezotte, 1993, p. 24). This definition allows for differences in achievement between groups, but requires equality in the distribution of achievement within groups.

The construction of test items and an instrument suitable for the assessment of very young children, i.e., kindergarten and prekindergarten, is quite problematic due to the wide range of physical, emotional, cultural and socioeconomic differences which are characteristic of preschool children (Dill, 1993; Kamii (Ed.), 1990; Shepard, Kagan, & Wurtz, 1998; Tuck, 1995; Wortham, 1995). For this reason Binet did not recommend assessing the intelligence of 7-year-old children by having them copy a written sentence or name coins, tasks which might be more reflective of socioeconomic status and cultural values than of mental capacity (Camilli & Shepard, 1994). Since the purpose of the Kindergarten Diagnostic Assessment was to determine what has been termed by Cummins (1981) as students' cognitive academic language proficiency (CALP) and basic interpersonal communications skills (BIC) rather than intelligence or mental capacity, the ability of the test items to discriminate between groups and thus target areas for instruction or instructional remediation was desirable. However, a well-designed assessment may also be useful for early interventions such as identifying precocious children (Javits 7+ Gifted and Talented Program: Kindergarten Assessment Model, Revised, 1991) or for diagnosing and preventing reading difficulties and assessing developmental stages of preschool children (Pflaum, 1974; Schweinhart, 1994; Shepard, Kagan, & Wurtz, 1998; Slavin & Madden,

1995; Snow, Burns, & Griffin, 1998; Snyder, 1993; Snyder & Sheehan, 1992a; Snyder & Sheehan, 1992b), particularly in school districts where there are large percentages of low income and limited English proficient families and their children (Gonzalez, 1996; Nurss, 1987; Zill, 1995). There are myriad confounding variables involved in examinations of the relationship between student achievement, ethnicity, and English proficiency and while it is beyond the scope of this paper to address them all a brief summary of some of the extant research is appropriate. According to U.S. Census data, non-Hispanic White children and Caribbean and African Black children are most likely to become acculturated to English within three years of entry into the United States, with Vietnamese and Mexican children requiring much longer to acquire English proficiency. Age, socioeconomic status, and previous educational opportunities have also been shown to play an important role in acculturation and acquisition of English (Rong & Preissle, 1998). Additionally there is considerable difference in the length of time that it takes for children to acquire survival English (i.e., basic communication skills) as compared to academic English. Most children will acquire basic interpersonal communications skills (BICS), also referred to as street or survival English during the course of informal interaction with their English speaking friends and acquaintances. The acquisition of academic language suitable for formal discourse is quite a different matter (Enright & McCloskey, 1988). The debate over how best to serve the educational needs of students of diverse national origins and ethnicities continues. However, there is a general consensus among linguists that (a) the length of time it takes for a child to learn a language is highly correlated to the child's native language, (b) there

is not a perfect correlation between the five or six common ethnic codes and native language[3], and (c) children and adults who have had previous opportunities to learn in their native language will acquire academic English more quickly (Clark, 1972; Enright & McCloskey; Krashen, 1982; Lisken-Gasparro, 1982; Lowe, 1982; Omaggio, 1986; Rong & Preissle; Sevilla, 1992). An assumption of the study was that students whose language in the home was other than English, who had recently immigrated into the United States, and who had not been provided with previous educational opportunities were classified as LEP and their scores were excluded from much of the analysis except where indicated.

The Kindergarten Diagnostic Assessment was designed by early childhood curriculum experts in an urban school district to include items that, according to accepted standards of early childhood education, would measure developmentally appropriate academic math and language skills for most preschool students (Engel, 1991; Wortham, 1995). According to data compiled from an NCES early childhood longitudinal study, in 1999 27% of prekindergarten female children were able to recognize all the letters of the alphabet compared to 21% of male children, 60% of female children were able to count to 20 or higher compared to 54% of male children. Twenty-nine percent of kindergarten children were able to recognize beginning sounds in the fall of their kindergarten year compared to 72% in the spring, with 2% recognizing "sight words" in the fall and 13% in the spring. Forty-three percent of female children (ages 3 - 5) demonstrated from three to four academic performance skills compared to 35% of male children. The skills of prekindergarten (age 3 -5) children who were identified as Hispanic were lower than those of

---

[3]According to the 1990 U.S. Census 3% of Asians, 9% of Hispanics, and 74% of White children are native born Americans, while of foreign born children living in the United States, 45% are Hispanic, 29% are Asian, and 16% are White. (Rong & Preissle, 1998).

children who were identified as either White or Black, with 60% of White children and 50% of Black children counting to 20 or higher compared to 41% of Hispanic children; 25% of White and Black children were able to recognize all the letters of the alphabet compared to 14% of Hispanic children (West, Denton, Germino-Hausken, 2000). For this reason there are recommendations to use alternative assessments as an aid to identifying culturally or non-verbally gifted Hispanic or other ethnic minority students with low English proficiency (Gonzalez, 1996; Scott, Duel, Jean-Françoise, & Urbano, 1996). According to the NCES study by the end of kindergarten gaps between the skills of male and female children had narrowed somewhat with female children slightly better at identifying number and shape and relative size and male children slightly better at ordinal sequence and addition/subtraction skills (West, Denton, Germino-Hausken). Another study found that over twice as many four-year-olds as three-year-olds were able to recognize most letters of the alphabet (Zill, 1995).

By the end of grade four NAEP data shows that 32% of male children and 28% of female children reach at least proficient level in mathematical estimation skills; by grade eight 22% of males and only 16% of females are proficient in estimation, while by grade twelve the gender gap widened with 47% of males and 31% of females proficient in mathematical estimation skills. The gap by ethnicity was even more pronounced, with 45% of White, 13% of Black, 18% of Hispanic, and 70% of Asian students at least proficient in estimation skills in grade twelve. However, females performed slightly better at grade eight than males in determining minimum measuring needed and in measuring lengths using a ruler (Mitchell, Hawkins, Stancavage, & Dossey, 1999). An examination of several sample questions for grade four estimation and probability from the 1996 NAEP summary data tables showed that male children did slightly better than female children on three multiple choice and performance questions and White

students did better than minority students on two of three sample questions, with American Indian students outperforming all other groups on one question identified by NAEP as hard. However, on three sample questions for grade eight mathematical estimation and prediction from the same 1996 NAEP table data females did better than males on each question.

Since the ability to perform the items on the Kindergarten Diagnostic Assessment had been determined by the school system to be essential to the demonstration of cognitive academic language proficiency (CALP) for kindergarten activities, an item by item examination served primarily as a means of comparison of kindergarten readiness across the system, rather than as a methodology for elimination or modification of test items. The language related items were consistent with items found on many early childhood assessments, such as ability to retell a story, associating letters and sounds, word recognition, book handling, following print on a page, making predictions about stories, and knowing letter sounds (Engel, 1991). Student test results were examined regularly since the school system served a mobile, majority-minority student population representing over 90 different native languages. Although simple group means can be useful in an initial examination they may not control for demographic variables of interest.

The assumption of unidimensionality, an underlying assumption of classical test theory, is also considered essential to the use of IRT as a method of determining item difficulty and student ability. Three assumptions basic to IRT are unidimensionality (UD), local independence (LI), and monotonicity (M); for dichotomously scored items M is assumed to hold true (Sijtsma & Hemker, 2000). Since the Kindergarten Diagnostic Assessment was designed to measure student readiness across several skill and performance areas the use of IRT for the test as a whole might be problematic, but could be useful to examine groups of items which measured either math or language ability. Unless traits measured by each item are readily apparent, it may

be necessary to use some method such as factor analysis or PCA to identify groups of items which measure the same or similar traits across the tested population. Also, it may be difficult to identify or separate unidimensionality in terms of its function of the tested items or the tested population. Thus, tests that are unidimensional for one population may not be so for another (Hambleton et al., 1991). This was the case with the present assessment as exploratory factor analysis resulted in different item clusters loading by subgroup. One item (item 5) on the assessment required that students be able to identify as well as draw a geometric figure, thus a correct response to the item represented a measure not only of cognitive ability but also of motor skills. However, there is much discussion in the literature that points out that test unidimensionality is a conceptual measure more often seen in theory than in practice.

Teachers, school administrators, and test coordinators may use a variety of approaches requiring more or less complexity of analysis to evaluate the effectiveness of the testing policies and programs in their school systems. These may range from relatively simple empirical graphical representations of item characteristics, to 2 x 2 contingency tables which compare χ^2 and effect sizes for performance of groups of interest on particular test items, ANOVA with post hoc effect size tests, computation and comparison of NCEs, logistic regression (Camilli & Shepard, 1994) and Δ and TID indices (Osterlind, 1983). However, these methods may not account for items which function differently at different ability levels for all the identified subgroups of interest, for example items that are more difficult for high achievers of one group may be easier for high achievers of another group. McLean and Lockwood (1996) provide an overview of the Angoff procedure and some of the basic issues involved in the scoring of norm-referenced, criterion-referenced, and performance assessments with an annotated bibliography of additional resources. A step-by-step overview of item analysis as it relates to individual

performance assessment is provided by Swezey (1981), who recommends the use of the ϕ statistic as a relatively conservative and simple method for computing reliability for dichotemous items. A review of the literature suggested that each of these approaches often yield somewhat different results and that the choice of most appropriate procedure is dependent upon a combination of considerations such as the purpose of the assessment, how it will be used, time constraints and available resources (Andrich, 1988; Camilli & Shepard, 1994; Cantrell, 1997; Lord, 1980; Snyder, 1993).

The Kuder Richardson scale (KR 20) is a commonly used scale which examines the internal consistency or reliability of a test. The higher the alpha coefficient, the more reliable the test. Alpha coefficients are computed for each item and KR 20 for the test for each subgroup of interest. The alpha coefficient then can be compared for each item by subgroup along with the total test reliability KR 20 for the subpopulations of interest. This method assumes that the items are not of equal difficulty.

The formula for computing KR 20 is:

$$r_{20} = \frac{n}{n-1}\left[1 - \frac{\sum pq}{s^2}\right]$$

where n is the number of test items, p is the percent passing the item, q is the percent missing the item and S^2 is the variance of the total scores. If the KR 20 for the total test is the same or nearly the same for each subpopulation tested, then the items of the test may be considered to measure the same skills across groups (Wiersma & Jurs, 1990). However, many test developers prefer to use additional methods to assure that tests do not discriminate against groups unfairly. The Golden Rule procedure looks for differences in p-values and recommends selection of items with

group differences of less than .15, however this may result in deletion of good items and loss of predictive validity of the test (Camilli & Shepard, 1994).

A fairly straightforward method for examining the differences between groups, when there are statistically significant differences between them, is to determine the effect size of the differences between the groups. This approach can be used not only for total test but also for individual items on the assessment, allowing administrators, counselors, teachers, and parents to make decisions about students' preparation and abilities based on greatest effect size differences, either positive or negative, between reference and focal groups, sometimes referred to as treatment and control or comparison groups. Even when there are statistically significant differences between groups the differences may not be large enough to be meaningful. While there are many variations on the effect size statistic the most commonly used and widely recognized effect size statistic is the standardized mean difference, or Cohen's d, with an effect size of at least .5 or larger being considered meaningful (not likely to have happened by chance) and an effect size of more than 1.00 being considered large (McLean, O'Neal, & Barnette, 2000; Shadish, Robinson, & Lu, 1999). For small groups this statistic can be computed with pencil and paper and for larger populations can be accomplished with a spreadsheet or statistical package. However, the specific statistic that is most appropriate would be determined by numbers of participants in each of the groups being compared and by range and variance of the scores of the participants in each of the groups (McLean et al.; Shadish et al.). Simply stated, Cohen's d is the difference between the mean of the two groups being compared divided by the pooled standard deviation of the groups, although the variations of this statistic that are most appropriate to use in each case are determined according to the considerations previously mentioned.

Transformed Item Difficulty (TID) or Δ scaling provides another relatively easy way to

examine item difficulties by converting item means to z scores and then to Δ scores using the formula:

$\Delta_i = 4z_i + 13$

where $_i$ is the item and z is the z score of the item mean. This method has been used by ETS among others as a way to place item difficulty on a common scale and to determine if the difference between groups for any item is greater than the average difference between groups on the test. The ETS method of using Δ to compute DIF involves comparison between scores of a reference group, usually White male students, and the focal groups of interest, usually females and minorities. The higher the Δ, the more difficult the item. Δs are usually ranked in order of difficulty and those for which the DIF exhibits the greatest difference between focal and reference groups are examined carefully to determine if the difference is due to item bias ("What's the DIF," 1999). In computing the DIF, a positive DIF means that the item was more difficult for the focal group and a negative DIF means the item was more difficult for the reference group. The term DIF is defined variously in the literature and there are numerous procedures for the computation but generally it is an umbrella term for any method used to identify possible test item bias and includes CT and χ^2 (Camilli & Shepard, 1994; Linn & Harnisch, 1981; Rudner et al., 1980). There are cautions for each of these methods as they may not differentiate between item difficulty, item discrimination, and DIF (Camilli & Shepard). In reviewing the literature it was apparent that over the course of time various researchers had used several of these terms interchangeably and therefore specification of the particular procedure to be used is advisable in the discussion of DIF, item difficulty, and item discrimination. The method for computing DIF will be determined, among other considerations, by the item type. Since the Kindergarten Diagnostic Assessment was based on student performance determined by

the teacher test administrator to be either mastery or nonmastery, the possible responses were coded as dichotemous responses rather than on a continuum or polytomous scale where performance scores are known to be less reliable than scores on multiple choice tests (Penfield & Lam, 2000).

Although some research suggests that the results of IRT and DIF may yield results similar to those provided by classical test calibration methods (Cantrell, 1997), the power of these methods lies in their ability to provide information about different types of test items ranging from dichotomous true-false type questions for nominal and ordinal test item responses to multiple choice responses. IRT (Item Response Theory) and DIF (Differential Item Functioning) provide models which are powerful in measuring item differences across focus groups and reference groups composed of respondents of varying ability levels for test items of varying levels of difficulty. In the classic three parameter IRT model, a student's ability (θ) is estimated (latent ability, the a parameter) by responses to items on a particular test given items of varying levels of discrimination or difficulty (the b parameter) after controlling for the probability that a student of relatively low ability was able to guess the correct answer to items of greater difficulty (the c parameter). Generally speaking, the use of the c parameter is more appropriate for examination of responses and items on a multiple choice test where respondents may guess than for performance items where guessing is less likely unless the performance items are scored on a polytomous scale.

The item characteristic curve (ICC) provides a graphical view which enables test developers to determine if individual test items or an entire test function differently for subpopulations of respondents. An ICC may be visually represented by plotting individual responses for comparison groups on a test item of interest to a graph (empirical ICC) or it may

be developed through a mathematical model known as a parametric ICC. Either of these methods provides a more accurate picture of differences in performance due to actual item bias than either ANOVA or simple comparison of point biserial correlation indices for two groups, as ANOVA procedures confound the constructs of item difficulty, item discrimination, and DIF while point biserials tend to be greatest for items of moderate difficulty (Camilli & Shepard, 1994).

The one-parameter Rasch model (1P) is the most popular of the IRT models. It provides a revised model for the a parameter indicative that the item does not discriminate between groups based solely on group membership but is able to discriminate between students of low or high ability on the construct being measured. In this model the b_i parameter is set where the probability of getting the correct response is .5 on the ability scale where a has the same value for all items on a test. Item characteristic curves generated using the 1P Rasch model are generally more nearly congruent than those using the two- or three-parameter model. (Camilli & Shepard, 1994; Hambleton et al., 1991). The equation for the 1P model is:

$P(\theta)=1/\{1 + \exp[-1.7a(\theta - b)]\}$.

If c parameter is zero under the assumption that students cannot or do not guess on the items or if examination of model fit indicates that the effect of guessing was negligible, the equation becomes a 2P model. The 2P model provides an item discrimination or b parameter to account for the ability of items to discriminate between respondents of high and low ability, meaning that students of higher ability will get the item correct and that the item will not be answered correctly by respondents who score low on the test. The point biserial correlation is most often used as a measure of the item discrimination parameter. Item discrimination indices may range from +1 to -1, with items having a point biserial correlation of less than .30 being

considered relatively poor indicators (Wiersma & Jurs). The equation for the 2P logistic model is:

$$P(\theta)=1/\{1 + \exp[-1.7\alpha(\theta - b)]\}.$$

Data and Methodology

The instrument was a 20 item assessment administered to 1137 prekindergarten students in an urban school district composed of 47% Black, 27% Hispanic, 19% White, and 7% other ethnic groups. Fifty-five percent of incoming kindergarten students were entitled to free or reduced price meals. The assessment was administered over a 15 day period in September to all incoming kindergarten students. Teachers read items to their students and recorded each response as a dichotomous variable, either correct or incorrect for each of the 20 items. Students were allowed a reasonable amount of time to respond as the test was not a timed test. Items assessed criteria such as student's ability to count out loud, identify and draw geometric shapes, retell, differentiate between real and make-believe, exhibit phonological awareness, hold a book correctly, and match voice to print (Appendix A). The assessment was administered only in English. Teachers recorded student responses on a matrix, information was then entered into the computer and matched by student ID number with demographic data including ethnicity, gender, SES, English proficiency, school, and DOB. Student ethnicity was encoded as either White, Black, Hispanic, Asian/Pacific Islander, or American Indian. For the logistic regression procedure ethnicity was recoded as minority (Black, Hispanic, Asian, American Indian) or nonminority (White). SES was encoded as either full, reduced, or free meal status and recoded to a dichotemous variable, (0 = full, 1 = free or reduced). Cases where missing data had been coded as 99 were recoded to system missing and did not enter the analysis. English proficiency was encoded as either LEP = 1 or LEP = 0. Categorical variables SES, gender, and ethnicity

that were used as covariates for the logistic regression were recoded as necessary and several procedures were used, including oversampling, to fit the model. SPSS 7.5 for Windows was used for data screening, t-tests, and for logistic regression. Excel was used to compute TID indices. Effect sizes were computed via statistical procedures available in SPSS and ES. Preliminary examination of descriptives included all cases and prior to the IRT analysis procedure, persons and items which either had no correct responses or all correct responses were identified. The IRT analysis was conducted using the Bigsteps program (Linacre & Wright, 2000) and figures were obtained by importing output data into SPSS and Excel spreadsheets. Scores for tested students who responded correctly to none of the items (63 non LEP students) and to all 20 of the items (14 non LEP students) did not enter this analysis as the items on the assessment could not measure the ability of these students. None of the 20 items was deleted from the IRT analysis as a result of this procedure as there were no items that were answered correctly by either all of the students or none of the students. While items can be calibrated such that they are constrained to the 1P model, the 2P model was used for the initial item and person fit analysis of the items on the Kindergarten Diagnostic Assessment. Because all items were read to the children and required an oral response such as naming characters in a story or performance of a task such as counting and moving blocks, the likelihood of guessing was slight. Responses were simply coded with a zero (criteria not mastered) or 1 (mastery of criteria). Although teachers had received inservice training in administering and scoring the Kindergarten Diagnostic Assessment to students each teacher made a determination as to whether student performance indicated mastery or non-mastery on an item.

The purpose of the present evaluation was to determine which, if any, of the differences were the result of item bias on individual test items and which were most likely the result of

opportunity to learn and thus were particularly well suited for identifying and assessing the instructional needs of identified groups of children. Had the Kindergarten Diagnostic Assessment results been used as a high stakes test rather than as a diagnostic tool, i.e., for admission into private school or special programs, there would, of course, have been much broader implications for test item bias (Childs, 1990; Meisels, 1987). There were significant differences in the socioeconomic status of the students by ethnicity, with Hispanic and Black students more likely to be entitled to free and reduced meal status than either White or Asian students.

Results

Group means, cumulative frequencies, norm curve equivalents (Appendix C), and coefficient alpha (KR 20) were computed for each of the subgroups of interest prior to further examination of individual test items. Several items for which gender or ethnicity appeared to be predictors of a correct response were then selected and results compared using ANOVA, contingency tables with chi square, effect size statistics for each item, TID, logistic regression, empirical ICC, and Rasch analysis by gender and ethnicity. Preliminary results (Table 1) showed that there were significant differences by identified subgroups on the assessment based on comparison of simple group mean scores for each item and percent correct for the total test for each subgroup. Group means for the 20 item assessment were not significantly different for male (n = 493, mean = 9.2) and female (n = 466, mean = 9.7) students, although t-tests indicated that three items were significantly more difficult for males than for females. Female prekindergarten children scored higher than male children on all but 3 items, including 5 items that measured math-related skills such as counting number of objects, using objects to tell a number story, comparing objects with non-standard units of measurement, and naming and drawing geometric

shapes. Female children scored significantly higher than male children on item 1 (one to one correspondence of items), item 4 (nonstandard measure units), and item 18 (understands that letters represent sounds). Item 19 (recognizes beginning consonant letters and sounds) was marginally significant. Male children scored higher than female children on 3 of the 20 items, making predictions, comprehension of stories, and matching voice to print, although not at a statistically significant level (Table 1).

Although there were statistically significant differences between male and female children on several of the items an examination of the effect size of the differences indicated that the effect of difference between male and female children on the items was small (item 1 $\underline{d}$ = .15, item 4 $\underline{d}$ = .12 and item 18 $\underline{d}$ = .13). Using Cohen's $\underline{d}$ resulted in a larger estimate of effect size for all three items than the use of the eta squared statistic.

In terms of facility or difficulty level of the three items where there was a significant difference between male and female children, item 1 was the least difficult item on the assessment (.8), item 4 had a difficulty level of .6, and item 18 of .2. Table 1 provides item difficulties and shows that of all the items on the assessment, items 16 and 6 were closest to the recommended ideal item difficulty level of .5, followed by item 4 and item 5 (Wiersma & Jurs, 1990).

Insert Table 1 about here

The mean raw score on the 20 item assessment was 9 items (SD = 4.9), for White students 12.1 (SD = 4.6), for Black students 8.8 (SD = 4.5), for Hispanic students 6 (SD = 4.7), and for Asian students 8 (SD = 5.0). There were statistically significant differences between ethnic groups on total test performance on the Kindergarten Diagnostic Assessment and an

examination of the effect size of group differences indicated a large effect for group membership with Hispanic students scoring, on average, approximately one and one third standard deviations below White students on the total assessment ($\underline{d}$ = 1.32) after excluding the scores of Hispanic students who were cross classified as also LEP. The effect size on the total assessment for Black students was .75 and for Asian students was .89 with White student mean score as the comparison, also after excluding the total mean scores of LEP students. A computation of effect size using group mean differences for items 16 and 6, the items that were in the .5 difficulty range, yielded a moderate to large effect size for comparison of White and Hispanic students (item 6, $\underline{d}$ = .62; item 16, $\underline{d}$ = .92).

Insert Figure 1 about here

Figure 1 shows that for minority and White students there was a greater difference between groups on the test than on the criterion. However these scores included those of limited English proficient (LEP) students who would be eligible for and receive ESL training based, in part, on their scores.

Insert Table 2 about here

Internal consistency as measured by KR 20 indicated that the assessment exhibited consistent psychometric properties across ethnic groups (White = .88, Black = .88, Hispanic = .91, Asian = .89) although group means were significantly different (Table 2). Individual items for which subgroup means and correlations were significantly different were selected for comparison with the following results. An examination of individual item alpha revealed that some items did have higher correlations than others depending upon group membership. Item 1, for example, was more highly correlated to total score on the KDA for Hispanic students (.63)

than for White students (.50), an indication that when graphed there would be a different item curve for the two groups on this item. In terms of item ability to discriminate between students of higher and lower ability, items 16, 17, 2, 4, and 5 were most powerful for predicting students who would have higher scores on the total assessment when using KR 20 coefficients.

Insert Table 3 about here

An examination and comparison of group means provided valuable information in terms of item facility or difficulty, items with a difficulty of .5 being considered ideal (Wiersma & Jurs, 1990). Additional analyses of test items were conducted to provide the most accurate group comparisons possible. These followup analyses yielded somewhat different results than comparison of simple group means or omnibus ANOVAs.

Logistic regression procedures were conducted with item 1 (Table 4), item 4 (Table 5), and item 18 (Table 6) respectively as dependent variables and SES, ethnicity, and gender as categorical variables. With ethnicity recoded for this procedure as a dichotemous categorical indicator, either minority or non-minority, results of the χ^2 test for each of these three items indicated that SES, gender and the ethnicity of the respondent were significant predictors of a correct response to each of the items with gender significant for each item after accounting for the effect of both ethnicity and SES (Tables 4 - 6).

Although the preliminary examination of the student scores for this data included scores of tested Limited English Proficient students, those student scores were excluded for the final analyses. After obtaining results for item 1, item 4, and item 18 using logistic regression, the items of the test were scaled using the Transformed Item Difficulty (TID) index. The Δ scale has a mean of 13 and a standard deviation of 4 and some authors recommend transforming the

scale to the more familiar 0 midpoint The resultant Δs and a DIF index for males and females are provided in Table 7. The items with larger values for Δ represent the more difficult items for the focal group of interest.

The Transformed Item Difficulty on Table 7 was ordered by difficulty according to gender, with the last column representing relative difficulty of items for male and female children after placing on the Δ scale. As can be seen, after placing on the Δ scale, 11 of the items were more difficulty for male children than for female children, but 8 items were more difficult for female children than for males. By standardizing the item difficulty (p-values) this procedure was able to correct for items that functioned differently by gender. Items 1, 4, and 18 were significantly more difficult for male children, although after DIF transformation males did better on items 2, 5, 7, 9, 14, 10, 11, and 12 than did females. There was no difference between males and females on the most difficult item on the assessment, item 13 (recognizes high frequency words in print). In fact, the greatest difference between male and female children after transformation to the TID index was on item 11 (makes predictions, males = 9.79, females = 10.69) and on item 12 (understanding of reading, verbalizes ideas, males = 15.84, females = 16.76). After standardizing the scale relative performance of male and female children on the test was more equally distributed.

Insert Table 7 about here

Graphing the TID data showed that item 10, an item of low or moderate difficulty, (differentiates between real and make-believe) and item 12 (verbalizes the central idea of a story) were more difficult for minority students (Table 7 and Figure 3, 4, and 5) than for White students. Item 10 proved to be the most inconsistent in difficulty in random samples drawn from

reference and focal groups of students. The greatest differences between focal and reference groups using the TID index were on item 10 (12.86, 8.59), item 12 (17.77, 14.51), item 13 (21.35, 18.22), and item 1 (9.04, 6.52). These differences were for Asian and Hispanic students compared to White students and therefore might be the result of differences in language or cultural comprehension. This is an indication of how much variability there is within groups of young children and points to the dangers of overinterpretation or overreliance on any one assessment instrument with prekindergarten children.

Using the logits for item difficulty obtained with the IRT analysis, empirical item characteristic curves were graphed against student scores for reference and focal groups to obtain test characteristic curves. A comparison of graphed models of TID and ICC shows that fitting the data to the IRT model provides a better fitting model with fewer items distant from the line of best fit. Use of the IRT model provided a generally higher total test $\underline{R}^2$ for each subgroup than did KR 20 or TID (males = .97, females = .97, White = .96, Black = .96, Hispanic = .96, Asian = .90).

Insert Table 8 about here

An examination of the a and b item parameter indices for the 2P model (Table 8) indicated that point biserial correlations (the a or item discrimination parameter) for all the items fell within the acceptable range from a low of .28 (item 13, a marginal item, recognizes high frequency words in print) to a high of .56 (item 18, understands that letters represent sounds) while the item difficulty or b parameters ranged from a high of 4.11 (item 13) to a low of -3.08 (item 9, demonstrates knowledge gained from pictures, one of the easier items). Fitting the Kindergarten Diagnostic data to the 2P IRT model would yield a preliminary scale score of 768

for White students who responded correctly to all 20 items, a scale score of 815 for Black students who responded correctly to all 20 items, a scale score of 852 for Hispanic students, and a scale score of 840 for Asian students who responded correctly to all 20 items on the Kindergarten Diagnostic Assessment although this is somewhat misleading as the scores of students who respond correctly to all items do not enter the analysis.

Insert Figures 6 through 12 about here

Test characteristic curves for each subpopulation are provided in Figures 6 through 12 with Figure 12 showing test curves for all subpopulations when constrained to the same scale for a preliminary comparison of groups. However, obtaining an optimal person- and item-fit may take multiple iterations and rescaling steps (Camilli & Shepard, 1994; Cantrell, 1997; Osterlind, 1983). A comprehensive analysis of test data, even using only one selected procedure, is very time-consuming and involves many steps in the identification of item-fit and person-fit. Thus time constraints and budget must be considered when selecting an analysis that will provide sufficient data for decision-making using available resources.

A comparison of the table data shows that items 13, 14, 18, 19, and 20 were consistently the most difficult for students, regardless of gender, ethnicity, or ability. In terms of the items' ability to predict students likely to do well on the assessment, item 16 (names uppercase letters out of sequence) had the highest correlation to total test score for all groups of students (.66), followed by item 17 (names lowercase letters out of sequence), and items 2, 4, and 5 (.64). In terms of IRT discriminant ability, item 18 had the highest point biserial correlation (.56) closely followed by items 16 and 17 (.54).

Tables and Figures showing item difficulty and scale scores are provided for each of the subgroups of interest in this study.

Conclusions

Although it is impossible to design and administer the perfect test or assessment, every effort was made to conform to best practices in program evaluation throughout the analysis and evaluation of this assessment. Specifically, consideration was given to following DEC guidelines regarding utility, feasibility, propriety, and technical adequacy of the report for its intended audience (Snyder, 1993).

The population on which this assessment was normed was not a representative population of prekindergarten students across the United States; it was, however representative of the population of many large urban school district prekindergarten populations. Therefore generalization of the results of this assessment to other populations of five and six year old children should be approached with caution. Standard conditions for administration of the kindergarten assessment were provided for through teacher in-service training. Teachers were also provided with a manual for assessment administration so that all children were read exactly the same instructions for each performance item and were scored to the same performance rubric for each item.

Ethnic identifications can be somewhat problematic (Schwartz, 1998). It was evident that at least in some cases the ethnic coding of children was questionable: all but two of the six children who were coded as American Indians had names such as Mohammed, Moshin, or Ahmad. Thus, in a school district where many children had parents who were not members of the same ethnic group, the coding for ethnicity was a forced choice code and there was no provision for indicating that a child was of mixed ethnicity. A 1996 NCES study using the FRSS

(Fast Response Survey System) with 926 respondents found that only 5% of schools surveyed used a category to indicate multirace students (Carey & Farris, 1996). The same study found that some school systems coded ethnicity only one time upon the child's entry into the system while other school systems asked the students to supply their ethnic identification at the beginning of each school year. The question that may well arise as a logical conclusion is whether the ethnicity of students should or should not be considered in the analysis of test scores. However, since educators must have some means of assuring that the unique educational needs of all groups of students are served it is likely that student scores will be disaggregated by ethnicity and gender for some time to come or at least for as long as there are significant differences between the groups.

School districts are placed in a rather problematic and tenuous position in the process of assessing preschool children. While the National Educational Goals Panel in its Principles and Recommendations for Early Childhood Assessments acknowledges the dual role that assessment plays in helping to identify mild learning disorders and in the monitoring of developmental trends in young children, it also cautions against the use of so-called high stakes tests for children 8 years of age and younger. Thus, while methods for evaluating items such as the Golden Rule procedure may have proven valuable under some circumstances, particularly for populations of adult test-takers who have completed much of their education and thus have presumably been provided with opportunities to learn throughout their years of schooling, it would not be appropriate for kindergarten students for several reasons. First, there is too much variability in cognitive and motor ability of children before age 8 to use assessments for purposes other than as a means to monitor trends in a child's development and to determine instructional needs (Shepard, Kagan, & Wurtz, 1998). The arbitrary grouping and placement of

children according to results of very early assessments can result, as one author so aptly observed, in children with low scores on cognitive academic language being consigned to the turtle group and spending the remainder of their elementary school years practicing drawing very straight lines from the bunny to its mother, deprived of the more challenging and rewarding activities assigned to their peers (Engel, 1991). It would be poor instructional practice and self-defeating to eliminate tasks that students were supposed to be able to perform from a performance assessment (Zill, 1995). Use of a guideline such as the Golden Rule (a difference in p-value of .15 for focal and reference group) would have potentially eliminated the best items, according to widely accepted test development standards, from the kindergarten assessment and would have left those that were so difficult that few students of any group answered them correctly or those that were so easy that all children responded to them correctly. At the time of the administration of the Kindergarten Diagnostic Assessment relatively few children (127) had been officially identified and coded as Limited English Proficient; however, the majority of tested children were Hispanic, Black, or Asian, and prior to final release of results an additional 141 students had been identified as LEP and recoded.

An important aspect of the use of this assessment was design of instruction for students from diverse socioeconomic and ethnic backgrounds so that they would begin school ready to learn while at the same time assuring that the tasks and vocabulary of the assessment were not biased or inappropriate for children of diverse ethnic heritage. Although it was not the purpose or within the scope of this analysis to evaluate the curricular goals and objectives which the items of the assessment measured, the tasks students performed were consistent with expectations for prekindergarten children (Kamii, 1990; Snow et al, 1998; Wortham, 1995).

Some of the results were not surprising. The effect of poverty on student achievement is well-known and there was a direct correlation between SES and ethnicity for this student population. Students who could count and who knew the alphabet were more likely to have higher scores on the assessment than those who could not perform those tasks.

According to the results of the assessment the gender gap for prekindergarten students was skewed in favor of female students, even on math-related items. However, NAEP assessments conducted in the spring of the kindergarten year as part of the Early Childhood Longitudinal Study and in later school years indicate that this gap closes rather rapidly, at least in mathematics (Mitchell et al., 1999). The Kindergarten Assessment did not include a math item that required children to make predictions, but male children did better than female children at making predictions about stories based on pictures or content. This finding is consistent with that of the ETS gender equity study (Cole, 1997) which found that on the average, male students outperformed female students on tests that measure mathematical estimation and prediction skills in high school and beyond.

Implications

There were statistically significant differences on at least some of the items between all subpopulations of interest in this evaluation, with the least difference between male and female children's performance and the greatest difference between White and Hispanic children. There are several implications for these findings.

1. Preschool female children did better than male children on nearly every knowledge and skill indicator on this assessment but, on the average, research findings show that they lose their slight advantage in math skills by high school or earlier. Do textbooks, curricular materials, and teachers use methods that are gender neutral or that do not encourage females to become or

remain interested in math and science?

2. Hispanic children did not do as well on the assessment as other identified ethnic minority populations. These students may have later qualified as LEP but not have been identified at the time of the assessment. It may also have been the result of these children not having access to learning opportunities in their native language or in their native country prior to entering kindergarten. Additionally, English is more difficult to learn for speakers of some languages than for others, depending upon the linguistic complexity of their native language and their prior exposure to native speakers of English.

3. It would be of interest to administer the assessment in both English and Spanish to Spanish-speaking Hispanic children to determine if their scores were the result of lack of knowledge of English or lack of prior opportunity to learn. This might not be a practical solution where there are populations of students from several different Spanish speaking countries and there is a great variation in local vocabulary and dialect from country to country.

4. Although children are expected to learn English and assimilate cultural mores as part of their educational experience, it is nonetheless true that in the cultures of some ethnic minorities different perceptions of time and of reality pertain, and that these cultural perceptions, independent of intelligence, ability, or prior opportunity to learn, may influence the ability of some students to perform well on assessments that ask them to differentiate between real and make-believe and to estimate measurements of time and space.

5. Although some items on the assessment were better than others in respect to difficulty indices and ability to discriminate between test takers of high and low ability, the internal consistency of the assessment was such that it would be useful for making decisions about groups of students.

6. Because the assessment was a performance assessment with students being required to perform multi-step tasks to demonstrate mastery of each item, it might be of interest to pilot a polytomous scoring scale or goal attainment scale so that students could be scored at partial mastery for items.

7. Particularly on performance assessments, children who are tested or evaluated by a familiar teacher have an advantage over those who are tested by a stranger, this is especially true for students who are bilingual or beginning English learners and who have become accustomed to the English speech patterns and pronunciation of their regular preschool teacher. When possible all children should assessed under like conditions, either by a familiar preschool teacher or by a stranger.

8. Because the items on this assessment were measures of attainment to a criteria or learning objective that all students would be expected to learn as a foundation for future success in school deletion of items that represented expected domains of attainment would not be advisable unless another item that measured attainment to the same domain or objective could be substituted.

9. There were substantial differences in results obtained depending upon the procedure used to calibrate items to the assessment and in the effect size of the differences between subpopulation performance on the assessment depending upon the specific effect size statistic that was used. This is often the result of the number and selection of items on the test, unequal ns for groups of participants being compared, or lack of homogeneity of variance within groups. The comparisons of the results obtained from use of several methods for examining each item of an assessment and the assessment as a whole may be helpful not only as a tool for improving instruction for students but also in future test item selection, item equating, and item banking.

Bibliography

Achilles, C. M. (1999). Let's put kids first, finally: Getting class size right. Thousand Oaks, CA: Corwin Press.

Andrich, D. (1988). Rasch models for measurement. Sage University series on Quantitative applications in the social sciences, 07-068. Newbury Park, CA: Sage Publications, Inc.

Bracey, G. & Rudner, L. M. (1992). Person-fit statistics: High potential and many unanswered questions. Practical Assessment, Research, & Evaluation, 3(7). Available online: http://ericae.net/pare/getvn.asp?v=3&n=7.

Camilli, G., & Shepard, L. (1994). MMSS: Methods for identifying biased test items. Thousand Oaks, CA: Sage Publications, Inc.

Cantrell, C. E. (1997). Item response theory: Understanding the one-parameter Rasch model. (ERIC Document Reproduction Service No. ED 415 281)

Carey, N., & Farris, E. (1996). Racial and ethnic classifications used by public schools. Statistical analysis report. (ERIC Document Reproduction Service No. ED 397 191)

Childs, R. A. (1990). Gender bias and fairness. Practical Research, Assessment, & Evaluation. Available online: http://ericae.net/pare/getvn.asp?v=2&n=3.

Clark, J. L. D. (1972). Foreign language testing: Theory and practice. Philadelphia: Center for Curriculum Development.

Cole, N. S. (1997). The ETS gender study: How males and females perform in educational settings. Princeton, NJ: Educational Testing Service.

Cummins, J. (1981). The role of primary language development in promoting educational success for language minority students. In Schooling and language minority

students: A theoretical framework. Los Angeles: Evaluation, Dissemination, and Assessment Center, California State University.

Dill, V. S. (1993). Closing the gap: acceleration vs. remediation and the impact of retention in grade on student achievement. The commissioner's critical issue analysis series, number 1. Texas Education Agency, Austin.

Engel, B. S. (1991). An approach to assessment in early literacy. In C. Kamii, M. Manning, & G. Manning (Eds.). Early literacy: A constructivist foundation for whole language. NEA Early Childhood Education Series. Washington, DC: National Education Association of the United States.

Enright, D. S., & McCloskey, M. L. (1988). Integrating English: Developing English language and literacy in the multilingual classroom. Reading, MA: Addison-Wesley Publishing Company.

Gonzalez, V. (1996) Identifying gifted bilingual Hispanic kindergartners with alternative sociocultural dual language assessments. Paper presented at the annual meeting of the American Educational Research Association (New York). (ERIC Document Reproduction Service No. ED 395 457)

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). MMSS: Fundamentals of item response theory. Newbury Park, CA: Sage Publications, Inc.

Javits 7+ Gifted and Talented Program: Kindergarten Assessment Model. Revised (1991). Community School District 18, Brooklyn, NY. (ERIC Document Reproduction Service No. ED 343 286)

Kamii, C. (Ed.). (1990). Achievement testing in the early grades: The games grown-ups play. Washington, DC: National Association for the Education of Young Children.

Kamii, C., Manning, M., & Manning, G. (1991). Early literacy: A constructivist foundation for whole language. NEA Early Childhood Education Series. Washington, DC: National Education Association of the United States.

Krashen, S. D. (1982). Principles and practices in second language acquisition. Oxford: Pergamon Press.

Lezotte, L. W. (1993). Creating effective schools today and tomorrow. Journal for Quality and Participation, 16(1), 22-30.

Linacre, J. M., & Wright, B. D. (2000). Bigsteps. [Computer program]. Chicago: Mesa Press.

Linn, R. L, & Harnisch, D. L. (1981). Interactions between item content and group membership on achievement test items. Journal of Educational Measurement, 18, 109-118.

Liskin-Gasparro, J. (1982). ETS oral proficiency testing manual. Princeton: Educational Testing Service.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Lowe, P. (1982). Manual for Language School oral interview workshops. Washington, D.C.: Defense Language Institute/Language School Joint Oral Interview Transfer Project.

McLean, J. E., & Lockwood, R. E. (1996). Why we assess students – and how: The competing measures of student performance. Thousand Oaks, CA: Corwin Press.

McLean, J. E., O'Neal, M. R., & Barnette, J. J. (2000, November). Are all effect sizes created equal? Paper presented at the annual meeting of the Mid-South Educational Research Association, Bowling Green, KY.

Meisels, S. J. (1987). High-stakes testing in kindergarten. Educational Leadership 46(7),

16-22.

Mitchell, J. H., Hawkins, E. F., Stancavage, F. B., & Dossey, J. A. (1999). Estimation skills, mathematics-in-context, and advanced skills in mathematics. National Center for Education Statistics. Washington, D.C.: U.S. Department of Education Office of Educational Research and Improvement.

Nurss, J. (1987). Readiness for kindergarten. ERIC Digest. (ERIC Document Reproduction Service No. ED 291 514)

Nye, B. (1993). Some questions and answers about multiage grouping. Paper presented at the Annual Conference on Creating Quality Schools. (ERIC Document Reproduction Service No. ED 384 998)

Omaggio, A. (1986). Teaching language in context. Boston: Heinle and Heinle Publishers, Inc.

Osterlind, S. J. (1983). Test item bias. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-030. Newbury Park: Sage Publications.

Penfield, R. D., & Lam, T. C. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. Educational Measurement: Issues and Practice, 19(3), 5-15.

Pflaum, S. W. (1974). The development of language and reading in the young child. Columbus, OH: Charles E. Merrill.

Rong, X. L., & Preissle, J. (1998). Educating immigrant students: What we need to know to meet the challenges. Thousand Oaks, CA: Corwin Press, Inc.

Rudner, L. M. (1994). Questions to ask when evaluating tests. Practical Assessment, Research, & Evaluation, 4(2). Available online: http://ericae.net/pare/getvn.asp?v=4&n=2.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). A Monte Carlo comparison of seven biased item detection techniques. Journal of Educational Measurement 17, 1-10.

Schwartz, W. (1998). The identity development of multiracial youth. ERIC/CUE Digest, Number 137. (ERIC Document Reproduction Service No. ED 425 248)

Schweinhart, L. J. (1994). Lasting benefits of preschool programs. ERIC Digest.,(ED 365 478)

Scott, M. S., Duel, L. S., Jean-Francois, B., & Urbano, R. C. (1996). Identifying cognitively gifted ethnic minority children. Gifted Children Quarterly, 40(3), 147-53.

Sevilla, J. (1992). Bilingual education: The last 25 years. In Anthony & Jacobson (Eds.). Helping at-risk students: What are the educational costs? (1992). Newbury Park: Sage Publications.

Shadish, W. R., Robinson, L., & Lu, C. (1999). ES: A computer program for effect size calculation. [Computer software]. St. Paul: Assessment Systems Corporation.

Shepard, L., Kagan, S. L, & Wurtz, E.(Eds.).(1998). Principles and recommendations for early childhood assessment. National Education Goals Panel. Washington, D. C.: U. S. Government Printing Office.

Sijtsma, K., & Hemker, B. T. (2000). A taxonomy of IRT models for ordering persons and items using simple sum scores. Journal of Educational and Behavioral Statistics, 25(4), 391-416.

Slavin, R. L., & Fashola, O. S. (1998). Show me the evidence! Proven and promising programs for America's schools. Thousand Oaks, CA: Corwin Press.

Slavin, R. L., & Madden, N. (1995). Effects of Success for All on the achievement of English Language Learners. Paper presented at the annual meeting of the American Educational

Research Association (San Francisco). (ERIC Document Service Reproduction No. ED 388 050)

Snow, C., Burns, S, & Griffin, P. (1998). Preventing reading difficulties in young children. Washington, D.C.: National Academy Press.

Snyder, S. (1993). Program evaluation: DEC Recommended Practices (ERIC Document Reproduction Service No. ED 370 267)

Snyder, S., & Sheehan, R. (1992). Rasch analysis of the standardization data of the Bayley Mental Scale of Infant Development. Diagnostique, 17, 185-94.

Snyder, S., & Sheehan, R. (1992). The Rasch measurement model: An introduction. Journal of Early Intervention, 16(1), 87-95.

Swezey, R. W. (1981). Individual performance assessment: An approach to criterion-referenced test development. Reston, VA: Reston Publishing Co., Inc.

Tuck, K. (1995). Assessment of the Continuous Progress Report for the early learning years (An examination of scale construction). (ERIC Document Reproduction Service No. ED 402 335)

West, J., Denton, K., & Germino-Hausken, E. (2000). America's kindergartners: Early childhood longitudinal study - Kindergarten class of 1998-1999. Report No. NCES 2000-070. National Center for Education Statistics. Washington, DC: U.S. Department of Education Office of Educational Research and Improvement.

What's the DIF: Helping to ensure test question fairness. (1999) Educational Testing Service: Author.

Wiersma, W., & Jurs, S. G. (1990). Educational measurement and testing. 2nd ed. Boston: Allyn and Bacon.

Wortham, S. C. (1995). Measurement and evaluation in early childhood education. 2nd ed. Englewood Cliffs, N. J.: Merrill.

Zill, N. (1995). School readiness and children's developmental status. ERIC Digest. (ERIC Document Reproduction Service No. ED 389 475)

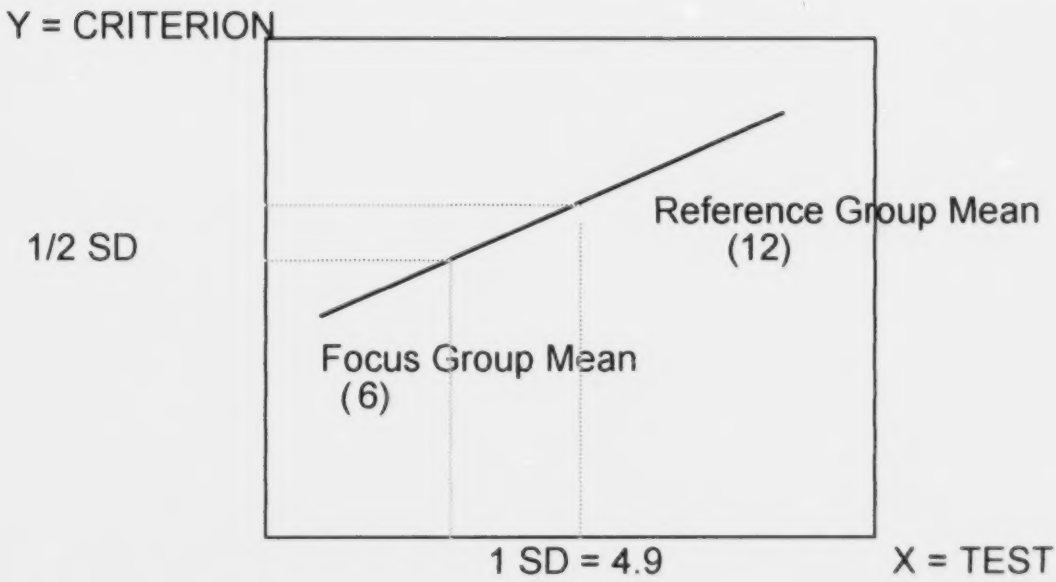

Figure 1. Comparison of two groups of children's mean scores based on Thorndike's concept of "performance fairness" if the difference between groups is greater on the test than on the criterion.

Table 1

T-Tests for Significant Difference in Item Performance by Gender, LEP Excluded
(Males = 495, Females = 467)

	Mean score				
Item	All	Males	Females	t	Sig t
1	81	78	84	-2.429	.015
2	74	73	74	-.623	.533
3	71	70	73	-.989	.323
4	61	58	64	-2.188	.029
5	41	40	42	-.627	.531
6	52	51	54	-1.139	.255
7	65	65	66	-.355	.723
8	76	74	78	-1.380	.168
9	86	85	86	-.374	.709
10	66	66	66	.307	.759
11	65	67	63	1.336	.182
12	26	29	24	1.655	.098
13	05	03	06	-1.269	.205
14	13	14	13	.314	.753
15	30	29	32	-.945	.345
16	46	45	48	-1.031	.303
17	33	31	35	-1.524	.128
18	18	15	20	-2.200	.028
19	16	14	18	-1.913	.056
20	18	16	20	-.623	.533

Table 2

Means by Ethnicity, LEP Excluded
(White = 207, Black = 522, Hispanic = 292, Asian = 63)

	Ethnicity							
Item	White	SD	Black	SD	Hispanic	SD	Asian	SD
1	.88	.33	.80	.40	.66	.47	.78	.42
2	.90	.31	.69	.47	.56	.50	.68	.47
3	.84	.37	.67	.47	.56	.50	.60	.49
4	.75	.44	.57	.50	.47	.50	.60	.49
5	.61	.49	.38	.49	.19	.39	.40	.49
6	.68	.47	.48	.50	.37	.49	.52	.50
7	.79	.41	.67	.47	.39	.49	.56	.50
8	.88	.33	.77	.42	.55	.50	.67	.48
9	.95	.21	.88	.33	.66	.47	.75	.44
10	.91	.28	.62	.49	.44	.50	.43	.50
11	.81	.39	.65	.48	.36	.48	.54	.50
12	.52	.50	.22	.41	.12	.33	.14	.35
13	.07	.26	.04	.19	.03	.16	.06	.25
14	.22	.41	.11	.30	.07	.26	.16	.37
15	.43	.50	.28	.45	.14	.34	.30	.46
16	.67	.47	.41	.49	.25	.43	.54	.50
17	.46	.50	.30	.46	.15	.36	.35	.48
18	.39	.49	.11	.32	.09	.29	.11	.32
19	.30	.46	.13	.33	.07	.26	.11	.32
20	.33	.47	.12	.33	.09	.29	.16	.37

Table 3

Item Correlation (KR 20) on Kindergarten Diagnostic Assessment

		Ethnicity				Gender	
Item	All	White	Black	Hispanic	Asian	Males	Females
1	.56	.50	.51	.63	.54	.47	.46
2	.64	.50	.60	.67	.70	.58	.50
3	.57	.50	.52	.59	.62	.45	.46
4	.57	.50	.54	.60	.51	.47	.45
5	.64	.60	.58	.62	.62	.50	.59
6	.57	.48	.58	.54	.58	.46	.49
7	.58	.47	.51	.59	.64	.50	.46
8	.56	.44	.49	.62	.44	.45	.46
9	.50	.39	.38	.57	.46	.42	.38
10	.55	.40	.48	.56	.50	.40	.45
11	.58	.49	.44	.64	.55	.48	.42
12	.52	.52	.46	.41	.48	.44	.46
13	.37	.38	.39	.38	.35	.29	.39
14	.49	.54	.45	.48	.54	.42	.44
15	.50	.37	.49	.54	.39	.37	.43
16	.66	.62	.64	.63	.54	.56	.59
17	.64	.69	.63	.57	.54	.53	.60
18	.62	.65	.57	.61	.61	.54	.61
19	.60	.62	.57	.61	.58	.51	.59
20	.55	.60	.50	.53	.59	.47	.52
TOTAL	.87	.88	.88	.91	.89	.87	.88

Table 4

Logistic Regression for Item 1 "Demonstrate One-to-One Correspondence Using Ten Items"

Variable	B	S.E.	Wald	df	R	Exp(B)
Meal status	.3715	.1869	3.9515	1	.0458*	1.4500
Ethnicity	.5727	.2639	4.7078	1	.0540*	1.7730
Gender	-.3915	.1688	5.3775	1	-.0603*	.6761

Note. LEP student scores excluded. All variables were entered in a single step. Parameter coding 1 = male, 0 = female, 1 = nonminority, 0 = minority, 1 = full price meal, 0 = free or reduced meal. Goodness-of-fit test $\chi^2(4) = .2261$, $p = .9941$.

* $p < .05$

Table 5

Logistic Regression for Item 4 "Uses Nonstandard Measure Units"

Variable	B	S.E.	Wald	df	R	Exp(B)
Meal status	.5877	.1505	15.2545	1	.1016***	1.7999
Ethnicity	.5314	.1975	7.2422	1	.0639**	1.7013
Gender	-.2969	.1360	4.7698	1	-.0464*	.7431

Note. LEP student scores excluded. All variables were entered in a single step. Goodness-of-fit test χ (4) = 1.1366, p = .8884

*** $p < .001$; ** $p < .01$, * $p < .05$

Table 6

Logistic Regression for Item 18 "Understands that Letters Represent Sounds"

Variable	B	S.E.	Wald	df	R	Exp(B)
Meal status	1.1607	.2138	29.4658	1	.1759***	3.1922
Ethnicity	1.0598	.2024	27.4225	1	.1692***	2.8859
Gender	-.4333	.1826	5.6304	1	-.0639*	.6484

Note. LEP student scores excluded. All variables were entered in a single step.Goodness-of-fit test χ (4) = 3.3580, p = .4998

*** $p < .001$, * $p < .05$

Table 7

Transformed Item Difficulty Index (TID) and DIF

ITEM	DELTAM	DELTAF	DELTAW	DELTAB	DELTAH	DELTAA	DIF M/F
1	8.09	7.42	9.04	7.71	6.52	6.94	-.66
4	11.19	10.53	11.02	11.14	10.05	9.99	-.65
18	17.85	17.38	16.49	17.99	17.10	18.28	-.47
8	8.71	8.36	9.04	8.16	8.56	8.80	-.35
17	15.37	15.05	15.42	15.16	15.99	14.22	-.33
20	17.70	17.38	17.40	17.84	17.10	17.43	-.32
19	18.01	17.69	17.85	17.69	17.48	18.28	-.32
3	9.33	9.13	9.65	9.65	8.38	9.99	-.19
6	12.27	12.09	12.08	12.48	11.90	11.34	-.18
16	13.20	13.02	12.23	13.52	14.13	11.00	-.18
15	15.68	15.67	15.88	15.46	16.18	15.06	-.01
13	19.55	19.56	21.35	19.03	18.22	19.46	.00
2	8.86	8.98	8.74	9.35	8.38	8.63	.12
5	13.82	13.96	13.14	13.97	15.25	13.37	.14
9	6.85	7.11	7.98	6.52	6.52	7.45	.26
7	10.10	10.38	10.41	9.65	11.53	10.66	.28
14	18.01	18.47	19.07	17.99	17.48	17.43	.46
10	9.79	10.38	8.59	10.39	10.60	12.86	.59
11	9.79	10.69	10.11	9.95	12.09	11.00	.90
12	15.84	16.76	14.51	16.35	16.55	17.77	.92

Figure 2

TID Correlation of Male and Female Students

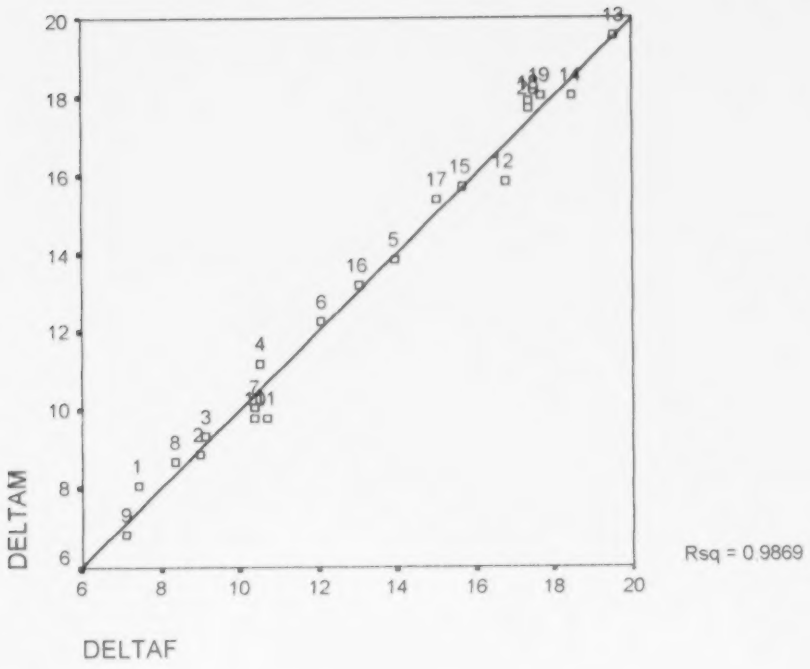

Figure 3

TID Correlation of White and Black Students

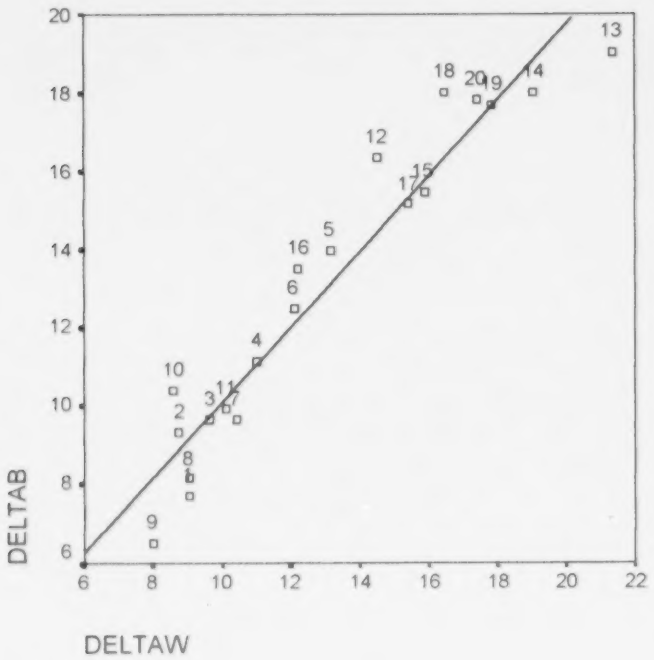

Figure 4

TID Correlation of White and Hispanic Students

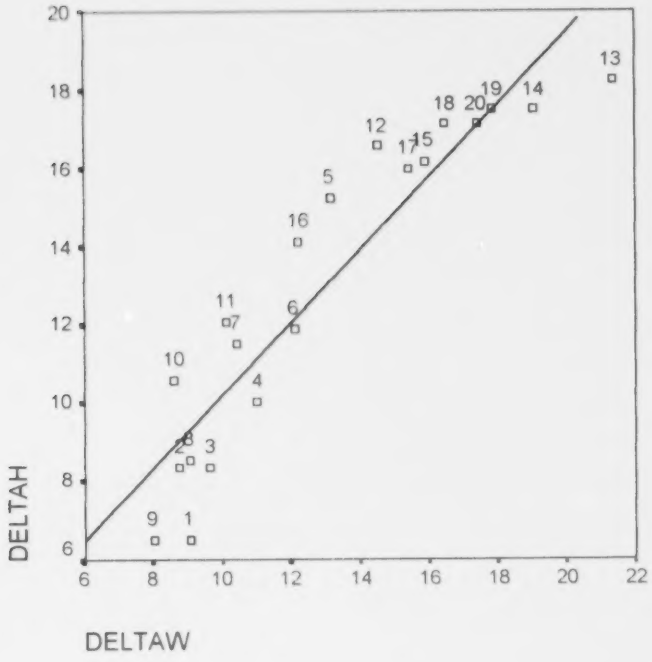

Figure 5

TID Correlation of White and Asian Students

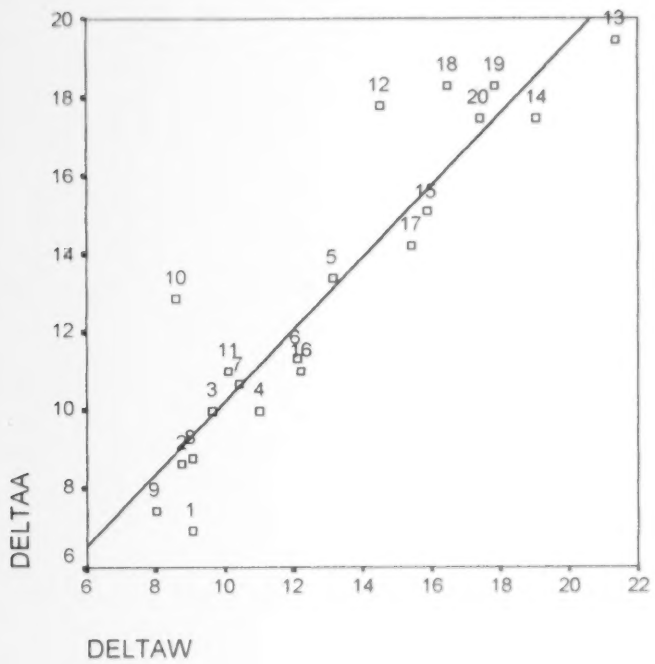

Figure 6

Item Characteristic Curve (ICC) Total Test for Male Students

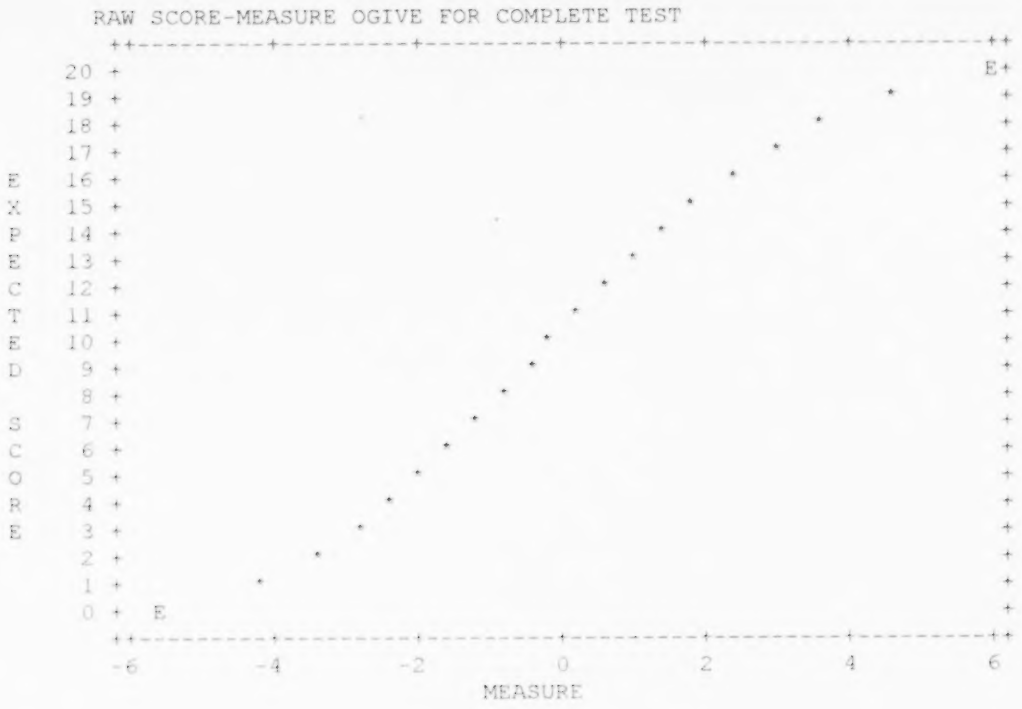

Figure 7

Item Characteristic Curve (ICC) Total Test for Female Students

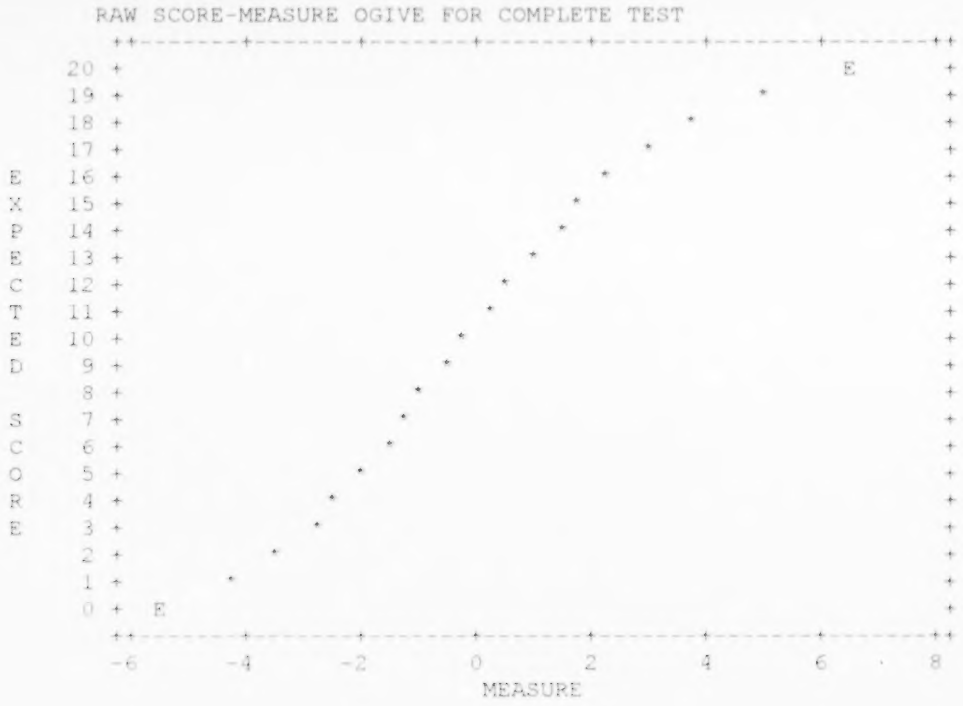

Figure 8

Item Characteristic Curve (ICC) Total Test for White Students

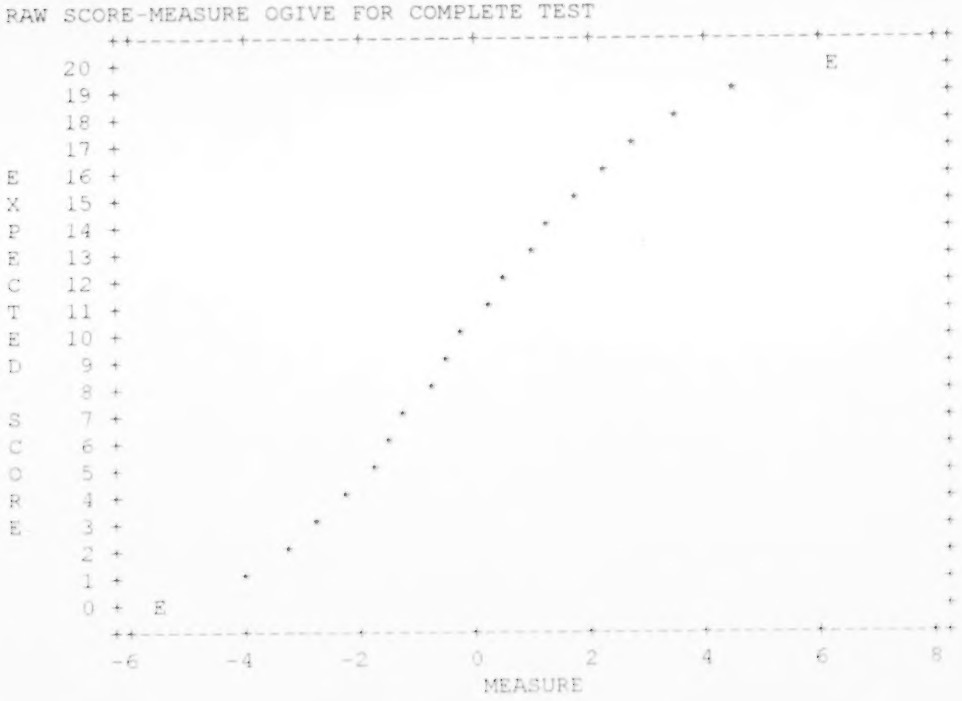

Figure 9

Item Characteristic Curve (ICC) Total Test for Black Students

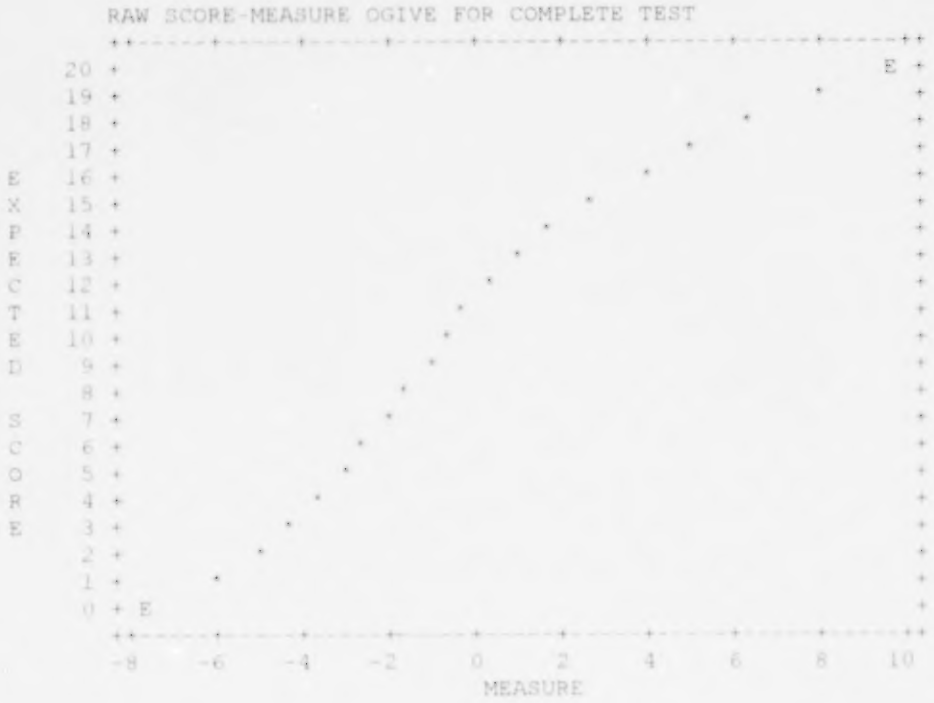

Figure 10

Item Characteristic Curve (ICC) Total Test for Hispanic Students

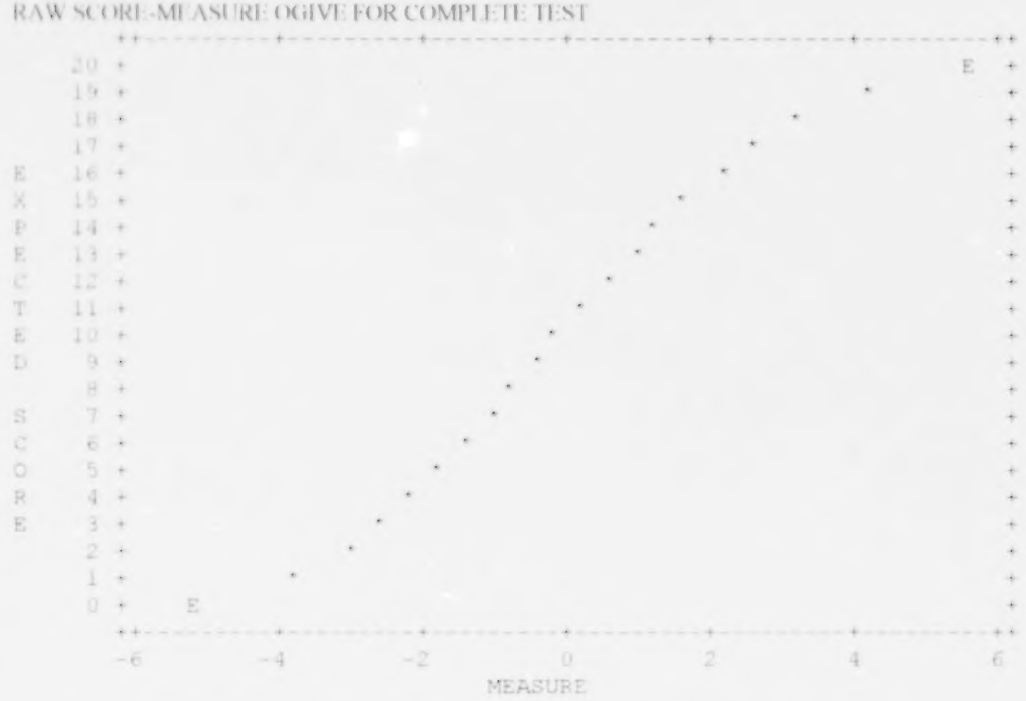

Figure 11

Item Characteristic Curve (ICC) Total Test for Asian Students

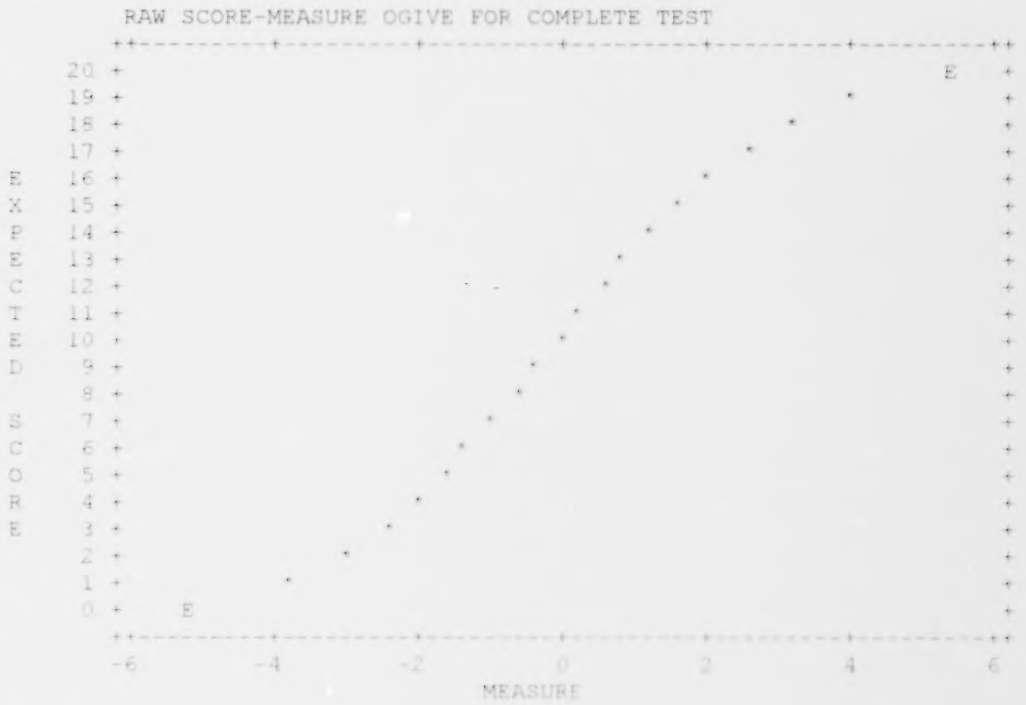

Figure 12

ICC Total Test for All Subgroups

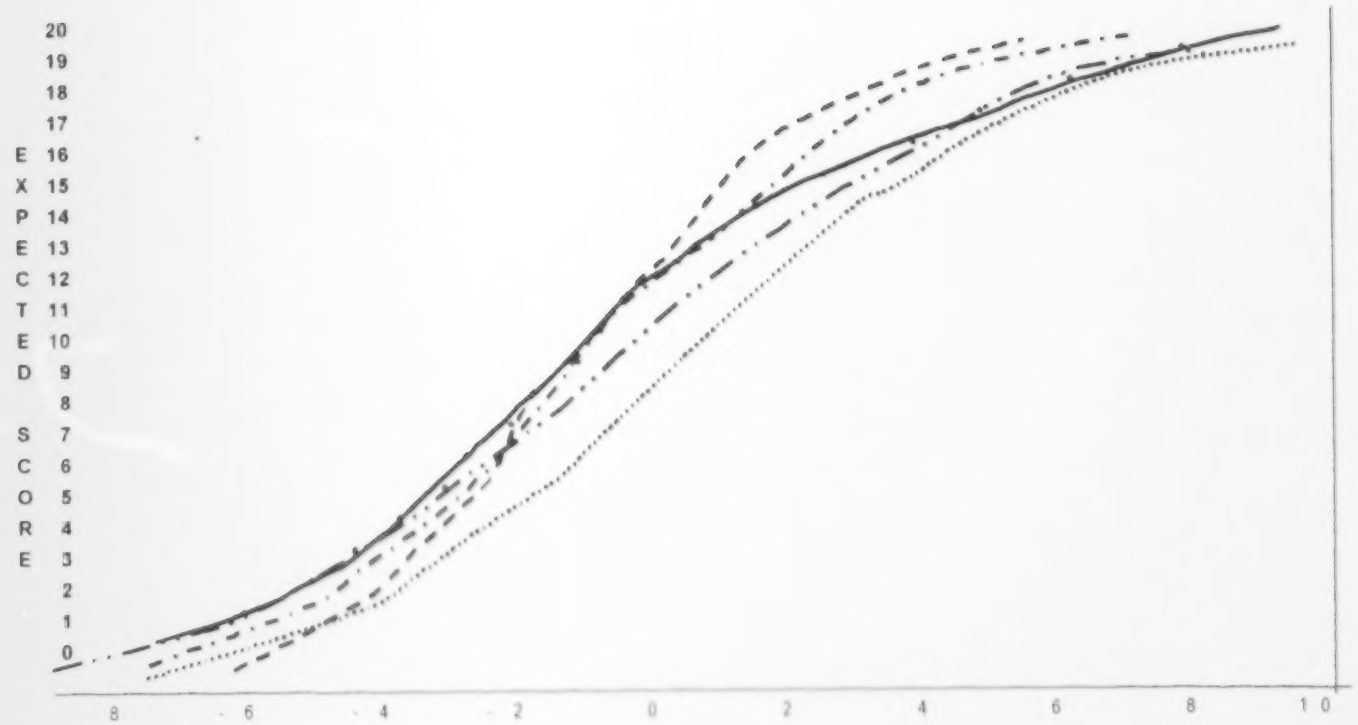

Table 8

Rasch Parameters by Gender and Ethnicity

Item	Males	Females	White	Black	Hispanic	Asian	All	PTBS
1	-2.39	-2.71	-2.04	-2.43	-2.73	-2.56	-2.39	0.39
2	-2.07	-2.08	-2.04	-1.53	-2.16	-1.74	-1.83	0.51
3	-2.07	-1.41	-1.32	-1.4	-2.16	-1.11	-1.63	0.41
4	-1.25	-1.62	-0.7	-0.81	-1.65	-1.23	-1.02	0.43
5	0.04	0.71	0.11	0.24	0.78	-0.12	0.21	0.51
6	-0.23	-0.87	-0.31	-0.54	-0.89	-0.77	-0.41	0.43
7	-1.32	-1.48	-0.86	-2.43	-1.04	-0.77	-1.31	0.43
8	-2.15	-2.25	-1.77	-2.71	-1.5	-1.35	-2.03	0.37
9	-3.04	-3.03	-2.98	-3.42	-2.43	-1.88	-3.06	0.29
10	-0.87	-1.00	-2.04	-1.04	-1.42	-0.34	-1.28	0.37
11	-1.45	-1.14	-1.32	-1.47	-1.11	-0.88	-1.31	0.39
12	1.53	1.08	0.82	2.05	1.24	1.86	1.19	0.39
13	3.80	5.14	4.55	4.95	3.97	3.16	4.11	0.28
14	2.39	3.14	2.97	2.4	2.45	1.68	2.46	0.40
15	0.49	0.39	0.95	0.71	1.67	0.44	0.98	0.35
16	0.34	-0.13	-0.31	-0.26	-0.28	-0.99	-0.06	0.54
17	1.01	1.18	1.27	0.41	1.12	0.33	0.82	0.54
18	2.39	2.06	1.2	2.69	2.02	2.29	2.13	0.56
19	3.22	2.63	2.3	2.54	2.45	2.29	2.33	0.53
20	1.66	1.39	1.53	2.05	1.67	1.68	2.09	0.47

NOTE. Measure computed from a random sample of each group except for measures for ALL and for PTBS (point biserial correlation) which were computed using score data for 836 non LEP kindergarten students.

APPENDIX A

Correlation Matrix for Kindergarten Diagnostic Assessment, NAEP, and Stanford Achievement Test (9) Objectives

Kindergarten Diagnostic Assessment Objective	Stanford Achievement Test Objective	NAEP
1. Demonstrate one to one correspondence using 10 items 2. Tell number of objects in group 3. Tell a number story	K.2. Given a set containing nine or fewer concrete items, tell how many are in the set by counting the number of items orally	(NCES: Count to 20 or higher)
4. Uses nonstandard measure units	K.12. Compare two objects or events, using direct comparisons or nonstandard units of measure, according to one or more of the following attributes: length, height, weight, temperature	G4. Recognize best measurement unit, know units of measure
5. Name and draw geometric shapes	K.13. Identify, describe, and make plane geometric figures (circle, triangle, square, and rectangle)	G4. Draw a geometric figure
6. Book handling	K.5. Understands how print is organized and read. Holds print materials in correct position. Identifies the front cover, back cover, and title page of a book. Follows words from left to right and from top to bottom on a printed page.	
7. Retell a story in own words	K.8. Demonstrates comprehension of stories	
8. Recognizes predictable patterns in text	K.8. Demonstrates comprehension of stories	
9. Demonstrates knowledge gained from pictures	K.8. Demonstrates comprehension of stories; verbalizes and/or demonstrates knowledge gained from pictures	

Kindergarten Diagnostic Assessment

Kindergarten Diagnostic Assessment Objective	Stanford Achievement Test Objective	NAEP
10. Distinguishes between real and make-believe	K.8. Demonstrates comprehension of stories	
11. Makes predictions	K.8. Demonstrates comprehension of stories; uses pictures and context to make predictions about story content prior to and during reading	
12. Understanding of reading; verbalizes ideas	K.8. Demonstrates comprehension of stories	
13. Identifies high frequency words in print		(NCES: "sight words")
14. Matches voice to print	K.5. Understands how print is organized and read. Holds print materials in the correct position. Identifies the front cover, back cover, and title page of a book. Follows words from left to right and from top to bottom on a printed page. Matches voice with print, associating oral phonemes, syllables, words, and phrases with their written form.	
15. Identifies author and illustrator	K.8. Demonstrates comprehension of stories. Identifies what an author and an illustrator do.	
16. Identifies uppercase letters	K.9. Identifies both upper and lowercase letters	(NCES: Identifies letters of the alphabet
17. Identifies lowercase letters	K.9. Identifies both uppercase and lowercase letters	(NCES: Identifies letters of the alphabet)
18. Understands that letters represent sounds	K.7. Develops an understanding of basic phonetic principles	
19. Recognizes beginning consonant letters, sounds	K.7. Develops an understanding of basic phonetic principles	(NCES: Beginning sounds)
20. Uses beginning consonants when writing	K.11. Draws pictures and/or uses letters and phonetically spelled works to write about experiences, stories, people, objects, or events.	

APPENDIX B

TABLE OF SAMPLE NORMS (500/100) AND FREQUENCIES CORRESPONDING TO COMPLETE TEST

SCORE	MEASURE	S.E.	NORMED	S.E.	FREQUENCY		CUM.FREQ.		PERCENTILE
0	-5.44E	1.35	204	106	1	1.0	1	1.0	1
1	-4.13	1.07	279	61	1	1.0	2	2.0	2
2	-3.28	.81	328	46	3	3.0	5	5.0	4
3	-2.71	.70	361	40	2	2.0	7	7.0	6
4	-2.26	.64	387	37	5	5.0	12	12.0	10
5	-1.87	.61	409	35	6	6.0	18	18.0	15
6	-1.50	.59	430	34	6	6.0	24	24.0	21
7	-1.16	.58	450	33	13	13.0	37	37.0	30
8	-.82	.58	469	33	9	9.0	46	46.0	42
9	-.48	.58	488	33	6	6.0	52	52.0	49
10	-.14	.59	508	34	9	9.0	61	61.0	57
11	.22	.60	528	34	7	7.0	68	68.0	65
12	.59	.61	550	35	8	8.0	76	76.0	72
13	.97	.63	572	36	8	8.0	84	84.0	80
14	1.39	.65	595	37	3	3.0	87	87.0	86
15	1.82	.67	621	38	2	2.0	89	89.0	88
16	2.30	.70	648	40	6	6.0	95	95.0	92
17	2.84	.76	679	44	1	1.0	96	96.0	96
18	3.50	.88	717	50	3	3.0	99	99.0	98
19	4.49	1.14	773	65	0	.0	99	99.0	99
20	5.82E	1.90	855	109	1	1.0	100	100.0	99

BEST COPY AVAILABLE

APPENDIX C

Summary Table of NCEs for Total Test by Gender and Ethnicity

	Mean NCE	N	Median
Males	51	521	51
Females	54	483	54
White	66	210	68
Black	52	518	52
Hispanic	42	218	43
Asian	51	54	52
Am Indian	50	4	48
TOTAL	52	1004	52